Contrasting Controlled Vocabulary and Tagging

Do Experts Choose the Right Names to Label the Wrong Things?

Paul Heymann and Hector Garcia-Molina
Department of Computer Science
Stanford University, Stanford, CA, USA
{heymann, hector}@cs.stanford.edu

ABSTRACT

Social cataloging sites—tagging systems where users tag books—provide us with a rare opportunity to contrast tags to other information organization systems. We contrast tags to a controlled vocabulary, the Library of Congress Subject Headings, which has been developed over several decades. We find that many of the keywords designated by tags and LCSH are similar or the same, but that usage of keywords by annotators is quite different.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*

General Terms

Experimentation, Human Factors, Measurement

1. INTRODUCTION

Collaborative tagging systems are one of the most common ways to organize data on the web. Sites like delicious, Flickr, YouTube, and others today each have millions of users and large tag ("keyword") vocabularies. However, to date, relatively little analysis has been done as to whether these tagging systems are effective at organizing data.

Recently, a new sort of tagging system has become common—social cataloging sites like LibraryThing where users tag books. Each book simultaneously has both expert assigned library metadata and user assigned tags. This situation presents for the first time an opportunity to compare expert created library data to tags from an uncontrolled vocabulary created by hundreds of thousands of users.

In previous work [4], we looked at library metadata as a basis for evaluating tagging systems as an information organization tool. We looked at questions like whether such systems could be federated and whether tags correspond to taxonomies like the Dewey Decimal Classification. Here, our focus is instead on the nature of keyword annotations, and specifically how user and expert keyword annotations differ. (The more recent experiments in this paper also apply a semantic relatedness measure in a novel way which we believe will be valuable more broadly in tagging systems.)

In this paper, we ask whether a controlled vocabulary of library keywords called the Library of Congress Subject Headings (LCSH) is different from the vocabulary developed by the users of LibraryThing. We find that many LCSH keywords correspond to tag keywords used by users of LibraryThing. However, we also find that even though an LCSH keyword and a tag may be syntactically the same, often the two keywords may annotate almost completely different groups of books. In our case, the experts seem to have picked the right keywords, but perhaps annotated them to the wrong books (from the users' perspectives). Thus, the common practice on the web of letting users organize their own data may be more appropriate.

2. PRELIMINARIES

A social tagging system consists of users $u \in U$, annotated keywords $a \in A$, and objects $o \in O$. We focus on social cataloging sites where the objects are books. More accurately, an object is a *work*, which represents one or more closely related books (e.g., one work can contain multiple editions).

An object o can be annotated with two types of keywords. If o is annotated by a site user, we call the keyword a *tag* (written $t_i \in T$). For example, a user can tag a work with "interesting" or "science fiction." At the same time, works are annotated with LCSH keywords (written $l_i \in L$) by librarians. For example, a work might be annotated with the LCSH keyword "Early Childhood Education."

In a given system, a keyword a implicitly defines a group, i.e., the group of all objects annotated with a (we define $O(a)$ to return this set of objects). A group also has a *size* equal to the number of objects it contains (we define $oc(a)$ to return this size). Since an object can have multiple annotations, it can belong to many groups. An object o becomes contained in group a when an annotator (either a user or a librarian) annotates o with a.

LCSH keywords come from a controlled vocabulary of hundreds of thousands of terms. Works are annotated with zero or more LCSH keywords. Each LCSH annotation consists of one "LCSH main topic" and zero or more "LCSH subtopics" selected from a vocabulary of phrases. For example, a book about the philosophy of religion might have keywords "Religion" (Main Topic) and "Philosophy" (Subtopic). We treat all main and subtopics as separate keywords.

WSDM '09 Barcelona, Spain

LCSH has some hierarchical structure. An LCSH keyword l_i has keywords which are "broader than" that keyword ($\{l_j, l_k \ldots\} \in B(l_i)$), "narrower than" that keyword ($\{l_j, l_k \ldots\} \in N(l_i)$), and "related to" the keyword ($\{l_j, l_k \ldots\} \in R(l_i)$). Unfortunately, this structure is not particularly consistent [1], in that if $l_k \in B(l_j)$ and $l_j \in B(l_i)$, it may not be the case that $l_k \in B(l_i)$. In practice, books rarely have more than three to six LCSH keywords due to originally being designed for card catalogs where space was at a premium. It is also common for only the most specific LCSH keywords to be annotated to a book, even if more general keywords apply. Lastly, because tags are annotated by regular users, and LCSH keywords are annotated by paid experts, $\{oc(l_j)|l_j \in L\}$ and $\{oc(t_i)|t_i \in T\}$ are quite different. Tags tend to focus on popular works, while keywords by paid experts annotate more works, less densely.

3. DATASET

Our source of library data is a dump of Library of Congress MARC records from the Internet Archive. We use only those $2,218,687$ records which had metadata we wanted for a variety of experiments. This required metadata included a Dewey Decimal Classification, a Library of Congress Classification, and an *ISBN* (a unique book identifier). When we refer to "LCSH keywords," we mean the value of MARC 650. MARC 650, strictly speaking, may include expert-assigned keywords from vocabularies other than LCSH, but in practice is made up almost entirely of that vocabulary in our dataset. Between April and October 2008, we crawled a sample of $309,071$ LibraryThing works based on a random selection of ISBNs from our Library of Congress data. Our analysis below looks only at works found in both LibraryThing and the Library of Congress, and only at the $8,783$ unique LCSH keywords and $47,957$ unique tags which annotate at least 10 works.

4. EXPERIMENTS

Our research question is, "how many keywords determined by expert consensus for LCSH are also used as tags, and are these keywords used in the same way?" In the experiments below, we divide this question as follows:

1. Section 4.1 asks whether LCSH keywords have syntactically equivalent tags. (For example, tag "java" is equivalent to LCSH "Java.")
2. Section 4.2 asks whether for a given syntactically equivalent (t_i, l_j) pair, t_i and l_j have the same prominence in lists ranked by $oc(t_i)$ and $oc(l_j)$.
3. Section 4.3 asks if syntactically equivalent (t_i, l_j) pairs are used in the same way by experts and users.
4. Section 4.4 asks whether LCSH keywords have semantically equivalent tags. (For example, "jewish life" is semantically equivalent to "jewish way of life.") We do not replicate the experiments from Sections 4.2 and 4.3 for semantic equivalence, but we expect less correlation and less similar usage between non-syntactically but semantically equivalent keyword pairs.

4.1 Syntactic Equivalence

Definition

The tag "painters" and the LCSH keyword "Painters" are obviously equivalent keywords. But is the tag "american science fiction" equivalent to "Science Fiction, American"? Is the tag "men in black" equivalent to "Men in Black (UFO Phenomenon)"?

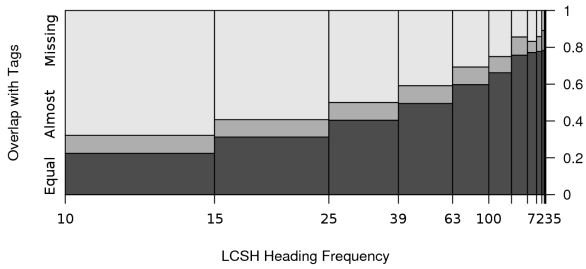


Figure 1: Spinogram showing probability of an LCSH keyword having a corresponding tag based on the frequency of the LCSH keyword. (Log-scale.)

We define two types of syntactic equivalence:

Exact The lower-cased tag is identical to the lower-cased LCSH keyword.

Almost Exact The lower-cased tag is identical to the lower-cased LCSH keyword if the LCSH keyword is modified to remove parenthetical remarks, swap the ordering of words around a comma, stem, or add or remove an "s."

Our "painters" example is exactly equivalent, while the other two examples are almost exactly equivalent. If there exists a tag t_i that is exactly or almost exactly syntactically equivalent to l_j, we say that $l_j \in S_{lcsh}$ and $(t_i, l_j) \in S_{pair}$.

Results

We found that $\frac{3408}{8783}$ LCSH keywords were exactly equivalent to a tag, while an additional $\frac{838}{8783}$ were almost exactly equivalent to a tag. In all, about 48% of LCSH keywords have equivalents according to one of the above two definitions. Such a high keyword overlap is all the more surprising given that many of the exactly equivalent LCSH keywords are multiple words, for example, "Vernacular Architecture" or "Quantum Field Theory."

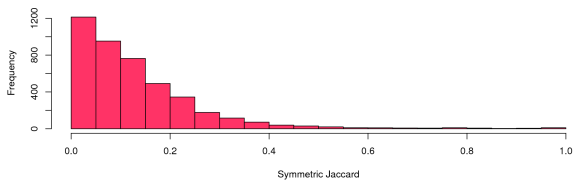
Cases where $l_j \notin S_{lcsh}$ are highly correlated with low $oc(l_j)$. Figure 1 shows the distribution of syntactic equivalence (y-axis) based on $oc(l_j)$ (x-axis). For example, if $10 \leq oc(l_j) \leq 15$, there is about a 30 percent chance that $l_j \in S_{lcsh}$ (and a 20 percent chance that l_j is exactly equivalent to some tag t_i). By contrast, if $63 \leq oc(l_j) \leq 100$, there is about a 70 percent chance that $l_j \in S_{lcsh}$. (We also suspect that longer LCSH keywords may be less likely to have syntactically equivalent tags because tags tend to be short.)

4.2 Rank Correlation of Syntactic Equivalents

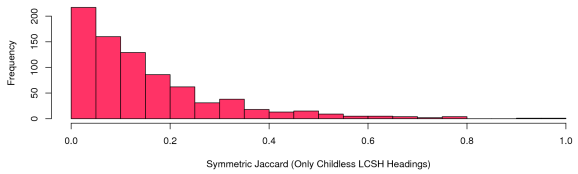
Are syntactically equivalent (t_i, l_j) pairs equally popular within their respective annotation types? For example, if the "java" tag annotates many works, does the "Java" LCSH keyword also annotate many works? We create two rankings of $\{(t_i, l_j) \in S_{pair}\}$, one ordered by $oc(t_i)$, the other ordered by $oc(l_j)$. We use Kendall's tau rank correlation to determine how similarly ranked the pairs are. For our data, $\tau \approx 0.305$. This means that the pairs are somewhat, but not highly, positively correlated. The experts and regular users have somewhat similar views of what the most important keywords are, but they do still differ substantially.

4.3 Expert/User Annotator Agreement

Do experts and regular users use the same keywords in the



(a) Histogram of J_{sym}



(b) Histogram of J_{sym}, $N(l_i) = \emptyset$

Figure 2: Symmetric Jaccard Similarity.

same ways? For example, many users in our dataset have annotated the book "The Wind in the Willows" with the tag "children's stories," yet no expert has annotated the book with the LCSH keyword "Children's Stories." We investigate the question of how common problems like these are below, and find that they are quite common.

Jaccard Similarities

We define three measures to try to get an idea of how commonly (t_i, l_j) pairs annotate the same books. We define symmetric Jaccard similarity as:

$$J_{sym} = \frac{|O(t_i) \cap O(l_j)|}{|O(t_i) \cup O(l_j)|}$$

For example, "children's stories" (above) has $J_{sym} = 0$, while "origami" has $J_{sym} = 0.75$. We also define two asymmetric Jaccard similarity measures, one for tags and one for LCSH:

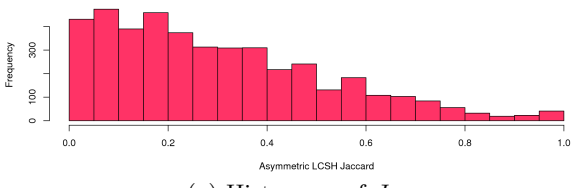
$$J_{tag}(t_i, l_j) = \frac{|O(t_i) \cap O(l_j)|}{|O(t_i)|} \quad J_{lcsh}(t_i, l_j) = \frac{|O(t_i) \cap O(l_j)|}{|O(l_j)|}$$

J_{sym} gives the ratio of the size of the intersection of two annotations to their union, so it may be dominated by one annotation if that annotation annotates many works. J_{tag} tells us what portion of the tagged works are covered by the LCSH keyword, and J_{lcsh} tells us what portion of LCSH annotated works are covered by the tag. For example, "knitting" has $J_{lcsh} = 0.97$ but $J_{sym} = 0.53$ because even though almost all works in $O(l_{knitting})$ are in $O(t_{knitting})$, $|O(t_{knitting})|$ is about twice as large as $|O(l_{knitting})|$.

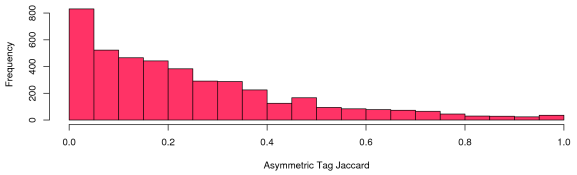
Results

For most $(t_i, l_j) \in S_{pair}$, $O(t_i) \cap O(l_j)$ is quite small. Figure 2(a) shows the distribution of J_{sym} for the $4,246$ (t_i, l_j) pairs in S_{pair}. The vast majority of such pairs have less than 20% overlap in work coverage.

A possible reason for small $O(t_i) \cap O(l_j)$ could be that librarians only choose the most specific appropriate LCSH keywords (see Section 2). In order to test this hypothesis, we computed J_{sym}, but only over LCSH keywords which were at the bottom of the LCSH hierarchy. In other words, we only chose l_i where $N(l_i) = \emptyset$. J_{sym} values for these pairs, shown in Figure 2(b) are very similar to those in Figure 2(a). This leads us to believe that specificity is not the core reason user and expert annotations differ.



(a) Histogram of J_{lcsh}



(b) Histogram of J_{tag}

Figure 3: Asymmetric Jaccard Similarity.

ESA	(t_i, l_j) pair
0.1	nature photography, indian baskets
0.2	fiction xxi c, angels in art
0.3	christian walk, women and peace
0.4	novecento/20th century, african american churches
0.5	20th century british literature, indians in literature
0.6	countries: italy, european economic community countries
0.7	medieval christianity, medieval, 500-1500
0.8	christian church, church work with the bereaved
0.9	detective and mystery fiction, detective and mystery stories

Table 1: Sampled (t_i, l_j) pairs with Wikipedia ESA values.

Figures 3(a) and 3(b) show the values of J_{lcsh} and J_{tag} for the $4,246$ pairs. Both show predominantly low Jaccard values. J_{lcsh} does have slightly higher Jaccard values, but it is still mostly below 0.4. A work labeled with an LCSH keyword is less than 50 percent likely to be labeled with the corresponding tag. A work labeled with a tag is even less likely to be labeled with the corresponding LCSH keyword.

4.4 Semantic Equivalence

Are there semantically, rather than syntactically equivalent tag/LCSH keyword pairs? In other words, are there many pairs like "middle ages" and "Middle Ages, 500-1500" where the meaning is the same, but the phrasing is slightly different? If so, how many?

Definition

We use semantic relatedness to determine whether (t_i, l_j) pairs are semantically equivalent. *Semantic relatedness* is a task where an algorithm gives a number between 0 and 1 for how related two words or phrases (w_1, w_2) are. For example, "vodka" and "gin" are highly related (closer to 1) while "rooster" and "voyage" are not (closer to 0). We use an algorithm called *Wikipedia Explicit Semantic Analysis (ESA)* [3] to calculate semantic relatedness. Wikipedia ESA calculates relatedness by looking at how often w_1 and w_2 co-occur in articles in Wikipedia. We write Wikipedia ESA as a function $sr_{esa}(t_i, l_j) \rightarrow [0, 1]$.

Understanding Wikipedia ESA Values

Table 1 shows representative Wikipedia ESA values for LCSH keywords $l_j \notin S_{lcsh}$. For example, for tag t_{ma}

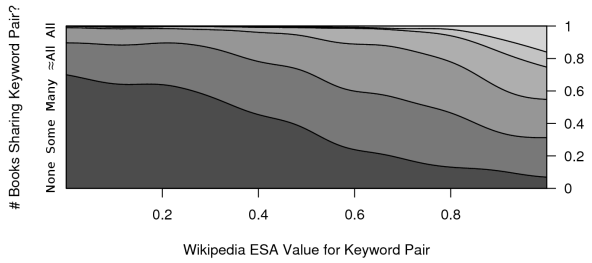


Figure 4: Conditional density plot showing probability of a (t_i, l_j) pair meaning that (t_i, l_j) could annotate $\{none, few, some, many, almostall, all\}$ of the same books according to human annotators based on Wikipedia ESA score of the pair.

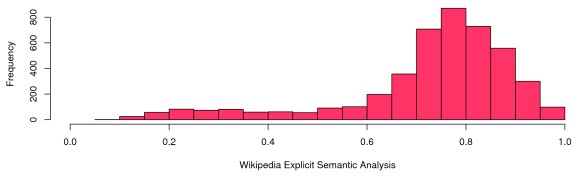


Figure 5: Histogram of Top Wikipedia ESA for Missing LCSH and All Tags.

"middle ages" and LCSH keyword l_{ma} "Middle Ages, 500-1500", $sr_{esa}(t_{dmf}, l_{dms}) \approx 0.98$ (not shown). By contrast, for t_{np} "nature photography" and l_{ib} "Indian Baskets", $sr_{esa}(t_{np}, l_{ib}) \approx 0.1$.

Figure 4 shows how Wikipedia ESA values translate into real relationships between (t_i, l_j) keyword pairs. We uniformly sampled (t_i, l_j) pairs where $l_j \notin S_{lcsh}$ by $sr_{esa}(t_i, l_j)$. We then asked human annotators how many books labeled with either t_i or l_j would be labeled with both t_i and l_j. Figure 4 shows sr_{esa} values on the x-axis and the distribution of answers $\in \{none, few, some, many, almostall, all\}$ on the y-axis. For example, at $sr_{esa} = 0.8$, 20 percent of keyword pairs have *many*, *almostall*, or *all* books in common (top three grays) according to human annotators. Likewise, more than half of pairs at $sr_{esa} = 0.8$ have at least *some* books in common by this measure. sr_{esa} is well correlated with how humans see the relationship between two keywords.

Results

We ran Wikipedia ESA over all (t_i, l_j) pairs where $l_j \notin S_{lcsh}$. Figure 5 shows $\{max(\{sr_{esa}(t_i, l_j) | t_i \in T\}) | l_j \in L - S_{lcsh}\}$. That figure shows that most of the non-syntactically equivalent LCSH keywords have a fairly semantically similar tag, with a Wikipedia ESA value between 0.7 and 0.9. By simulation using the probabilities from Figure 4, we estimate that $\approx$ 21 percent of $l_j \notin S_{lcsh}$ have a tag matching *all* or *almostall* of the keyword and $\approx$ 56 percent have a tag matching *many* books annotated with the keyword.

5. DISCUSSION

We looked at how a mature controlled vocabulary built over decades by experts contrasts with an uncontrolled vocabulary developed by hundreds of thousands of users over a few years. We found many (about 50 percent) of the keywords in the controlled vocabulary are in the uncontrolled vocabulary, especially more annotated keywords. We also found using a semantic relatedness measure that most of the remaining LCSH keywords have similar, though not exactly equivalent, tags. This suggests that often the keywords selected as controlled vocabulary keywords are the keywords that users naturally use to describe works.[1]

However, we found little agreement as to how to apply shared keywords. Sets of works annotated by corresponding LCSH keywords and tags rarely intersect significantly. This is true even if we merely check whether a corresponding tag annotates most of the works annotated by an LCSH keyword. This suggests one of three interesting possibilities:

1. Users and experts use many of the same keywords, but ultimately differ heavily as to how to apply them.
2. Experts are not allowed, or do not have time, to annotate works with all of the appropriate keywords.
3. Experts only label highly representative works with a term, rather than all works that might be considered to have the term, leading to low recall.

All of these possibilities are ultimately bad for retrieval using expert assigned controlled vocabularies.

When users and experts differ in how they annotate objects, we believe it is most reasonable to defer to the users. To say otherwise would be, in essence, to tell users that they do not know how to organize their own collections of objects. Ultimately, given that keywords are used by the users for navigation and browsing, we should evaluate the usefulness of annotations from their perspective, rather than the perspective of experts.

Our work also suggests an interesting alternative view on the vocabulary problem [2], a long standing observation in the world of human-computer interaction. The vocabulary problem suggests that given an object, people will choose many different names for that object. However, our work suggests that given a name (a tag in our case), people, whether experts or not, may disagree substantially on what objects that name should annotate.

6. REFERENCES

[1] M. Dykstra. LC Subject Headings Disguised as a Thesaurus. *Library Journal*, 113(4):42–46, 1988.

[2] G. Furnas, T. Landauer, L. Gomez, and S. Dumais. The Vocabulary Problem in Human-System Communication. *Comm. ACM*, 30(11):964–971, 1987.

[3] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *IJCAI '07*.

[4] P. Heymann and H. Garcia-Molina. Can Tagging Organize Human Knowledge? Technical report. Available: `http://ilpubs.stanford.edu/878/`.

[1]Keywords can be in one of three groups, but we focus on $L \cap T$ and $L - T$ in this paper, ignoring $T - L$. In our previous work [4], we found that about half of the 47,957 tags $t_i \in T$ are likely to be non-objective, non-content tags like "funny," "tbr," or "jiowef." We suspect that the balance of $T - L$ that is not syntactically equivalent to LCSH keywords is still either related to the LCSH keywords or describe completely different (objective, content-based) concepts.